NUS National University of Singapore | IMS Institute for Mathematical Sciences

SCIENTIFIC REPORTS

Efficient Sampling Algorithms for Complex Models

14 Jul 2025–25 Jul 2025

Organizing Committee

Tiangang Cui
The University of Sydney

Xin Tong
National University of Singapore

Lester Mackey
Stanford University

Alexandre Hoang Thiery
National University of Singapore

Jing Dong
Columbia University

Qiang Liu
The University of Texas at Austin

CONTENTS PAGE

TOWARD BALLISTIC ACCELERATION FOR LOG-CONCAVE SAMPLING

SINHO CHEWI

Classification AMS 2020: 60H35, 65C05, 65C40

Keywords: Harnack inequality, hypocoercivity, log-concave sampling, shifted composition, space-time Poincaré inequality, underdamped Langevin diffusion

This talk is based on a joint work with J. M. Altschuler and M. S. Zhang [4]. We consider the problem of sampling from a target density $\pi \propto \exp(-V)$ over $\mathbb{R}^d$, given access to gradient evaluations of V. A standard approach to this problem is to discretize the Langevin diffusion, which is the solution to the stochastic differential equation $\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$. If $\nabla^2 V \succeq \alpha I \succ 0$, then it is classical that the law μ_t of X_t converges toward the stationary distribution π with a quantitative decay rate: $\chi^2(\mu_t \,\|\, \pi) \le \exp(-2\alpha t)\,\chi^2(\mu_0 \,\|\, \pi)$. Algorithmic implementation requires discretization, and if we additionally impose the smoothness assumption $\nabla^2 V \preceq \beta I$, then the continuous-time decay rate is expected to yield algorithms for log-concave sampling with iteration complexities scaling as $O(\kappa)$, where $\kappa := \beta/\alpha$ is the condition number of the problem. Obtaining algorithms which achieve this expected rate, as well as enjoying good scalings with the ambient dimension and the target accuracy, has been the subject of intensive research in the past decade [6].

What if we want to converge faster? The underdamped (or kinetic) Langevin dynamics augments the state space with a momentum variable, leading to the SDE system $\mathrm{d}X_t = P_t\,\mathrm{d}t$, $\mathrm{d}P_t = \{-\nabla V(X_t) - \gamma P_t\}\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t$, where $\gamma > 0$ is a friction coefficient. Recently, these dynamics have been shown to achieve the "accelerated" convergence rate $\chi^2(\mu_t \,\|\, \pi) \lesssim \exp(-\Omega(\sqrt{\alpha}\,t))\,\chi^2(\mu_0 \,\|\, \pi)$, via a novel space-time Poincaré inequality [5]. For algorithmic implementations, this is expected to yield algorithms with a dependence of $O(\sqrt{\kappa})$ on the condition number, analogously to the acceleration phenomenon in convex optimization. However, prior works on discretization analysis were too lossy to obtain any result with $o(\kappa)$ dependence.

In our work, we build upon the shifted composition framework—introduced in our prior works [1, 2, 3]—to address this problem. Our main result states that randomized midpoint discretization, together with the space-time Poincaré inequality, leads to a sampling algorithm with condition number dependence $\widetilde{O}(\kappa^{5/6})$. The main challenge here is that, compared to the simpler Langevin dynamics, adaptation of shifted composition to the underdamped Langevin dynamics is considerably more challenging due to the hypoelliptic nature of the latter process.

REFERENCES

[1] J. M. Altschuler, S. Chewi. Shifted composition I: Harnack and reverse transport inequalities. 2024. *IEEE Transactions on Information Theory*, 1–1.

[2] J. M. Altschuler, S. Chewi. Shifted composition II: shift Harnack inequalities and curvature upper bounds. 2023. arXiv preprint 2401.00071.

[3] J. M. Altschuler, S. Chewi. Shifted composition III: local error framework for KL divergence. 2024. arXiv preprint 2412.17997.
[4] J. M. Altschuler, S. Chewi, M. S. Zhang. Shifted composition IV: underdamped Langevin and numerical discretizations with partial acceleration. 2025. arXiv preprint 2506.23062.
[5] Y. Cao, J. Lu, L. Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. 2023. *Arch. Ration. Mech. Anal.*, **247**(5), 90.
[6] S. Chewi. *Log-concave sampling*. 2025+. Draft available online at `https://chewisinho.github.io`.

YALE UNIVERSITY
Email address: `sinho.chewi@yale.edu`

DEEP TENSOR TRAIN APPROXIMATION OF TRANSPORT MAPS FOR BAYESIAN INVERSE PROBLEMS

TIANGANG CUI, SERGEY DOLGOV, ROBERT SCHEICHL

Classification AMS 2020: 65D15, 65D32, 65C05, 65C40, 65C60, 62F15, 15A69, 15A23, 65N21, 65L09

Keywords: Rare events, Bayesian inference, inverse problems, tensor train, transport maps

1. INTRODUCTION

Estimating expectations of random variables is central to uncertainty quantification, but rare events—occurring with very small probabilities—pose significant challenges. Standard Monte Carlo methods are inefficient because such events are seldom observed. Importance sampling (IS) mitigates this by biasing the sampling distribution toward rare events, yet designing effective importance distributions in high dimensions remains difficult, especially for concentrated or multimodal densities.

We address this problem in the context of high-dimensional Bayesian inverse problems, where expectations are taken with respect to posterior distributions conditioned on data [1]. We introduce a *deep importance sampling* framework that combines functional tensor-train (TT) decompositions with the deep inverse Rosenblatt transport (IRT). The method constructs importance distributions as compositions of TT-based maps, adaptively approximating the optimal IS density. We provide a theoretical analysis of variance and bias, and demonstrate scalability and accuracy on differential equation models involving extremely small probabilities.

2. BACKGROUND AND PROBLEM SETUP

Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ be a random variable with prior density π_0. The objective is to estimate the expectation $F = E_{\pi_0}[f(X)]$ for a function f. In rare event estimation, $f(x) = \mathbb{1}_{\mathcal{A}}(h(x))$ is an indicator function, where h is a response function and $\mathcal{A}$ is a failure set. The probability of failure is then $\mathrm{pr}_{\pi_0}(h(X) \in \mathcal{A})$.

In the Bayesian setting, given data y, the posterior density is $\pi^y(x) = \mathcal{L}^y(x)\pi_0(x)/Z$, where $\mathcal{L}^y$ is the likelihood and Z is the normalizing constant. The posterior failure probability is $E_{\pi^y}[f(X)]$.

The optimal IS density for estimating F is $p^*(x) \propto |f(x)|\pi_0(x)$. For posterior expectations, the ratio estimator

$$\hat{R} = \frac{\hat{Q}}{\hat{Z}}, \quad \hat{Q} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(x^i)\mathcal{L}^y(x^i)\pi_0(x^i)}{p(x^i)}, \quad \hat{Z} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathcal{L}^y(x^i)\pi_0(x^i)}{q(x^i)}$$

is used, with optimal densities $p^* \propto |f|\mathcal{L}^y\pi_0$ and $q^* = \pi^y$. The challenge is to approximate the densities $p \approx p^*$ and $q \approx q^*$ and their normalizing constants $\hat{Q} \approx Q, \hat{Z} \approx Z$ accurately in high dimensions.

3. Deep Importance Sampling with Tensor Trains

The core of the proposed method is the approximation of the optimal IS density using a deep composition of transformations. Each transformation is built via a squared TT decomposition of the square root of an unnormalized density.

3.1. **Squared Inverse Rosenblatt Transport.** We approximate the square root of the unnormalized optimal density $\rho^*(x) = |f(x)|\pi_0(x)$ (or its posterior counterparts) using a functional TT decomposition:

$$\sqrt{\rho^*}(x) \approx \tilde{g}(x) = \mathbf{G}_1(x_1) \cdots \mathbf{G}_d(x_d).$$

This leads to an approximate density:

$$p(x) = \frac{1}{\zeta} \left(\tilde{g}(x)^2 + \tau\lambda(x) \right),$$

where λ is a reference density (e.g., the prior), ζ is the normalizing constant, and $\tau > 0$ ensures $\mathrm{supp}(p^*) \subseteq \mathrm{supp}(p)$. The Hellinger distance between p and p^* is controlled by the TT approximation error.

The IRT $\mathcal{Q}$ is then constructed such that the pushforward of λ under $\mathcal{Q}$ equals p. This map is lower-triangular and evaluated via a sequence of one-dimensional conditional distribution functions, enabling efficient sampling.

3.2. **Deep Composition of Maps.** For rare events, the optimal IS density may be highly concentrated. To address this, we propose a deep composition of maps:

$$\mathcal{T}^{(L)} = \mathcal{Q}^{(1)} \circ \cdots \circ \mathcal{Q}^{(L)}.$$

Each layer $\mathcal{Q}^{(\ell)}$ is built to push forward a reference density to approximate an intermediate density $p^{(\ell)}$ (or $q^{(\ell)}$ for the posterior), which gradually approaches the target ρ^* (or π^y). This layered approach adapts to complex density structures.

The algorithm proceeds by recursively applying the squared IRT construction to the pullback of the intermediate densities under the current composite map. The final density $\bar{p} = \mathcal{T}^{(L)}_\sharp \lambda$ is used for IS. The estimator for the normalizing constant is:

$$\hat{\zeta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\rho^*(\mathcal{T}^{(L)}(U^i))}{\bar{p}(\mathcal{T}^{(L)}(U^i))}, \quad U^i \sim \lambda.$$

3.3. **Theoretical Analysis.** We provide a detailed analysis of the estimator's properties. Under mild assumptions, the estimator is unbiased and its variance is bounded by the Hellinger distance between the true and approximate IS densities. Key lemmas establish:

Lemma 3.1. [1, Lemma 3.6] *The relative variance* $\mathrm{var}_{\bar{p}}(p^*/\bar{p})$ *satisfies*

$$\mathrm{var}_{\bar{p}}(p^*/\bar{p}) \leq C_p D_H(p^*, \bar{p}),$$

or, under stronger assumptions,

$$\mathrm{var}_{\bar{p}}(p^*/\bar{p}) \leq C_m D_H(p^*, \bar{p})^2.$$

For the ratio estimator used in posterior expectations, we show:

Lemma 3.2. [1, Lemmas 3.8, 3.9] *The ratio estimator* $\hat{R}$ *is asymptotically unbiased, and its asymptotic variance is minimized when the samples used for the numerator and denominator are positively correlated.*

4. Application to Rare Event Estimation

The method is specialized for rare event estimation by smoothing the indicator function $f(x) = \mathtt{1}_{\mathcal{A}}(h(x))$ using a sigmoid function:

$$f_\gamma(x) = \left[1 + \exp\left(\gamma(a - h(x))\right)\right]^{-1},$$

which converges to the indicator as $\gamma \to \infty$. The sequence of intermediate densities uses increasing γ values to gradually sharpen the approximation.

For posterior rare event probabilities, we employ likelihood tempering for the denominator $q^{(\ell)} \propto (\pi^y)^{\beta_\ell}$ and combined smoothing and tempering for the numerator $p^{(\ell)} \propto f_{\gamma_\ell}(\mathcal{L}^y \pi_0)^{\beta_\ell}$. This allows the method to handle the challenges of both rare events and unnormalized posteriors.

5. Numerical Experiments

We present extensive numerical experiments on two models: a spatial SIR model and a groundwater contaminant transport model.

5.1. **Spatial SIR Model.** The SIR model describes the spread of an infectious disease through a network of compartments. The goal is to estimate the posterior probability that the number of infected individuals in a compartment exceeds a threshold. The problem dimension is $d = 2K$, where K is the number of compartments.

Results show:

- The method scales linearly with dimension d (number of parameters).
- The Hellinger error increases only moderately as the event probability decreases.
- The ratio estimator benefits from positive correlation between numerator and denominator samples.
- The method outperforms the cross-entropy method, which fails for $K \geq 3$ even with large sample sizes.

5.2. **Groundwater Contaminant Transport.** The model involves a PDE describing groundwater flow and an ODE for contaminant transport. The rare event is the breakthrough time of a contaminant being below a threshold. The diffusivity field is uncertain and represented via a Karhunen–Loève expansion.

Key findings:

- The posterior risk can differ significantly from the prior risk, highlighting the importance of using data.
- The method accurately estimates probabilities as low as 10^{-17}.
- The method again significantly outperforms the cross-entropy method in high dimensions.

6. Related Work

The proposed method connects several areas of research:

- **Importance Sampling and Rare Events.** Traditional IS methods often use parametric families (e.g., Gaussian mixtures) for the biasing distribution [5, 4]. These can be inefficient in high dimensions. The cross-entropy method [4] adaptively fits a parametric family but struggles with complex, high-dimensional densities.

- **Functional Tensor Decompositions** [3, 9] provide a scalable way to approximate multivariate functions. The use of TT in Bayesian inference includes [6, 7]. The IRT [11] has been used in variational inference [12, 2]. The novel contribution here is the *squared* IRT and its deep composition for IS.
- **Deep Generative Models.** The deep composition of maps is inspired by deep generative models. However, instead of training a neural network, we use TT-cross approximation, which offers faster computations.
- **Multilevel and Multifidelity Methods.** Other approaches for rare events include multilevel Monte Carlo [8, 13] and multifidelity methods [10]. These can potentially be combined with the proposed method for further efficiency.

REFERENCES

[1] Tiangang Cui, Sergey Dolgov, and Robert Scheichl. Deep importance sampling using tensor trains with application to a priori and a posteriori rare events. *SIAM Journal on Scientific Computing*, 46(1):C1–C29, 2024.

[2] Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics*, 24:2063–2108, 2024.

[3] Daniele Bigoni, Allan P. Engsig-Karup, and Youssef M. Marzouk. Spectral tensor-train decomposition. *SIAM Journal on Scientific Computing*, 38(4):A2405–A2439, 2016.

[4] Zdravko I. Botev and Dirk P. Kroese. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability*, 10(4):471–505, 2008.

[5] Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

[6] Sergey Dolgov, K. Anaya-Izquierdo, Colin Fox, and Robert Scheichl. Approximation and sampling of multivariate probability distributions in the tensor train decomposition. *Statistics and Computing*, 30:603–625, 2020.

[7] Martin Eigel, Robert Gruhlke, and Manuel Marschall. Low-rank tensor reconstruction of concentrated densities with application to Bayesian inversion. *Statistics and Computing*, 32(2):1–27, 2022.

[8] Daniel Elfverson, Fredrik Hellman, and Axel Målqvist. A multilevel Monte Carlo method for computing failure probabilities. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):312–330, 2016.

[9] Ivan V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[10] Benjamin Peherstorfer, Tiangang Cui, Youssef Marzouk, and Karen Willcox. Multifidelity importance sampling. *Computer Methods in Applied Mechanics and Engineering*, 300:490–509, 2016.

[11] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[12] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[13] Fabian Wagner, Jonas Latz, Iason Papaioannou, and Elisabeth Ullmann. Multilevel sequential importance sampling for rare event estimation. *SIAM Journal on Scientific Computing*, 42(4):A2062–A2087, 2020.

UNIVERSITY OF BATH
Email address: s.dolgov@bath.ac.uk

AUTOMATED PDMP SAMPLING AND SCALING LIMITS

KENGO KAMATANI

AMS 2020 Classification: 60J25, 65C05, 65C40, 62F15.
Keywords: piecewise-deterministic Markov process (PDMP), Bouncy Particle Sampler, Zig-Zag, Forward Event-Chain, thinning, adaptive horizon, grid envelope, scaling limits.

Piecewise-deterministic Markov processes (PDMPs) provide nonreversible, rejection-free Monte Carlo samplers that expected to mix faster than reversible MCMC on complex targets. The talk is divivded into two parts. First, it describes an automated way to simulate PDMP event times by thinning with a tight, piecewise-constant envelope built on a time grid and combined with an adaptive window proposed in [2]. Second, it summarises scaling-limit guidance for choosing among Bouncy Particle (BPS), Zig-Zag (ZZS), and Forward Event-Chain (FEC) samplers. A Python/JAX implementation is available at `https://github.com/charlyandral/pdmp_jax`.

Let $\Pi(\mathrm{d}x) \propto e^{-U(x)}\mathrm{d}x$ be the target and consider the lifted process $Z_t = (X_t, V_t)$ with deterministic flow between random events and velocity updates at events. Standard intensities are $\lambda_{\mathrm{BPS}}(x,v) = \langle \nabla U(x), v\rangle_+ + \underline{\lambda}$ and $\lambda_{\mathrm{ZZ}}(x,v) = \sum_{i=1}^d (v_i\,\partial_i U(x))_+$. Exact event times solve $\int_0^\tau \lambda(X_s,V_s)\,\mathrm{d}s = E$ with $E \sim \mathrm{Exp}(1)$; in practice one simulates by thinning against an envelope $\Lambda \geq \lambda$ on a local time window.

Envelope construction on a grid, following [2], is straightforward and computationally efficient. Fix a horizon $t_{\max}$ and a partition $0 = t_0 < t_1 < \cdots < t_N = t_{\max}$. Evaluate λ and its time derivative λ' at all grid points. On each segment $[t_i, t_{i+1}]$ set a constant bound

$$\Lambda_i = \max\{\lambda(t_i), \lambda(t_{i+1}), m_i\},$$

where m_i is the ordinate of the intersection of the endpoint tangents. Define $\Lambda(t) = \Lambda_i$ for $t \in [t_i, t_{i+1})$. Because Λ is piecewise constant, its integral is piecewise linear, so solving $\int_0^\tau \Lambda = \mathrm{Exp}(1)$ reduces to a simple running sum across segments. Two modest design choices tend to improve robustness in practice: for BPS/FEC, build the envelope using the *signed* inner product $\langle \nabla U(X_t), V_t\rangle$ and apply the positive part only at the end; for ZZS, construct per-coordinate envelopes (optionally in signed form) and sum them. This mitigates the local loss of derivative information introduced by the $[\cdot]_+$ operation and helps maintain reasonably tight bounds.

The horizon adapts automatically. If the process frequently advances to $t_{\max}$ without an accepted event, enlarge the window ($t_{\max} \leftarrow \alpha_+ t_{\max}$); if thinning rejections accumulate, shrink it ($t_{\max} \leftarrow t_{\max}/\alpha_-$). Gentle factors such as $\alpha_+ \approx 1.01$ and $\alpha_- \approx 1.04$ stabilise the trade-off between frequent re-bounding (too small) and low acceptance (too large). Vectorised evaluation on the grid typically outpaces per-window optimisation (e.g. Brent) while using all function/derivative values to shape the envelope.

Correctness follows from a mild separation condition. Let Z be the union of local maxima of λ and zeros of λ'' on $[0, t_{\max}]$, and let δ be the minimum distance between distinct points of Z. If λ is C^2 and the grid mesh is strictly smaller than δ, then $\Lambda(t) \geq \lambda(t)$ for all t, hence thinning with Λ is exact. Intuitively, each segment contains at most one

local maximum or change of concavity, so the graph of λ lies below the two endpoint tangents. In implementation, a diagnostic ensures safety: if a proposed time T ever yields $\lambda(T)/\Lambda(T) > 1$, halve $t_{\max}$, refine the grid, and redraw.

Computation is dominated by gradient evaluations needed to build Λ and to compute accept ratios at proposals; the deterministic flow is nearly free. Writing N_T for the number of accepted events in $[0, T]$ and M_T for the number of windows processed, one has $\mathbb{E}[N_T] = \mathbb{E}\int_0^T \lambda_t \, \mathrm{d}t$. Under a stabilised envelope ($c_1 \leq \mathbb{E}\int_\Delta \Lambda_t \, \mathrm{d}t \leq C_1$ and $\inf_{t\in\Delta} \lambda_t/\Lambda_t \geq c_2$ on each window Δ), this yields $\mathbb{E}[N_T] \asymp \mathbb{E}[M_T]$, so the expected gradient count is proportional to $\mathbb{E}[N_T]$. Empirically, on a two-Gaussian mixture with a sharp secondary mode, the grid-based envelope with adaptive horizon attains the correct mean near $(0.5, 0.5)$ while running markedly faster than Brent-based maximisation. On a 20-mode local mixture, the vectorised-signed ZZS bound drastically reduces both the frequency and the size of envelope violations compared with non-vectorised bounds; for BPS, the signed strategy eliminates violations with modest grids.

Scaling limits offer concise guidance. In high dimension on approximately isotropic targets, ZZS mixes in $O(1)$ event times (under unit-speed normalisation) with $O(d)$ jumps per unit time, giving total work $O(d)$; BPS requires $O(d)$ mixing with $O(d)$ jumps, totalling $O(d^2)$ [5]. Under strong anisotropy with a stiff direction of scale ε, BPS maintains $O(1)$ mixing and $O(\varepsilon^{-1})$ jump rates (total $O(\varepsilon^{-1})$), whereas ZZS needs $O(\varepsilon^{-1})$ mixing and $O(\varepsilon^{-1})$ jump rates (total $O(\varepsilon^{-2})$), favouring BPS [6]. During burn-in from poor initialisation, a fluid-limit analysis suggests jump-rate scalings of $O(\varepsilon^{-1/4})$ (BPS) and $O(\varepsilon^{-1/2})$ (ZZS), while an FEC variant keeps $O(1)$ rates, making FEC attractive for rapid contraction toward the typical set [1].

REFERENCES

[1] Agrawal, S., Bierkens, J., Kamatani, K. and Roberts, G. O. (2025+), Transient Regime of Piecewise Deterministic Monte Carlo Algorithms (in preparation)

[2] Andral, C. and Kamatani, K. (2024). Automated techniques for efficient sampling of PDMPs. *arXiv:2408.03682*.

[3] Andrieu, C., et al. (2021). Hypocoercivity of Piecewise Deterministic Markov Process-Monte Carlo. *Ann. Appl. Probab.*

[4] Bierkens, J., Fearnhead, P. and Roberts, G. O. (2019). The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data. *Ann. Statist.*

[5] Bierkens, J., Kamatani, K. and Roberts, G. O. (2022). High-dimensional scaling limits of piecewise deterministic sampling algorithms. *Ann. Appl. Probab.*

[6] Bierkens, J., Kamatani, K. and Roberts, G. O. (2025). Scaling of Piecewise Deterministic Monte Carlo for Anisotropic Targets. *Bernoulli*, to appear.

[7] Bouchard-Côté, A., Vollmer, S. and Doucet, A. (2018). The Bouncy Particle Sampler: A Nonreversible Rejection-Free Markov Chain Monte Carlo Method. *JASA*.

[8] Corbella, A., Spencer, S. and Roberts, G. O. (2022). Automatic Zig-Zag sampling in practice. *Statistics & Computing*.

[9] Davis, M. H. A. (1984). Piecewise-deterministic Markov processes. *J. Royal Stat. Soc. B*.

[10] Michel, M., Durmus, A. and Sénécal, S. (2020). Forward Event-Chain Monte Carlo. *J. Comput. Graph. Stat.*

[11] Sutton, M. and Fearnhead, P. (2023). Concave-convex PDMP-based sampling. *J. Comput. Graph. Stat.*

THE INSTITUTE OF STATISTICAL MATHEMATICS, TOKYO, JAPAN
Email address: kamatani@ism.ac.jp

PROVABLE GUARANTEES FOR SAMPLING MULTIMODAL DISTRIBUTIONS

HOLDEN LEE

Classification AMS 2020: 60J25, 68W20

Keywords: sampling, multimodal distribution, Markov chain Monte Carlo, simulated tempering, sequential Monte Carlo

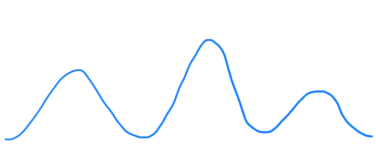
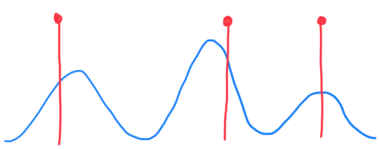
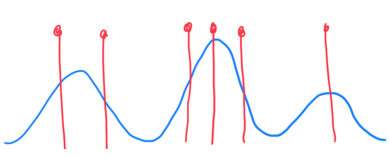
[1] Multimodal distributions pose significant challenges for sampling algorithms, because local algorithms can easily get stuck in a single mode. A variety of methods inspired by statistical physics have been developed to address this problem, but with limited theoretical understanding. We show provable guarantees for sampling under three settings of increasing information:

(1) No advice: only access to the unnormalized density
(2) Weak advice: with warm start points to each of the modes
(3) Strong advice: with a few samples from the distribution

For (1), we give conditions under which simulated tempering [GLR18] and sequential Monte Carlo are effective [LS24]. For (2), we show an algorithm based on Annealed Leap Point Sampling (ALPS) can sample under generic conditions, including unbalanced cases not covered by (1) [LS25]. Finally, for (3), we show efficient sampling *without changing the Markov chain* with a number of data samples almost-linear in the number of modes [KLV25], with applications to score-based models and pseudolikelihood estimation.

[1]Slides are available at `https://www.dropbox.com/scl/fi/lmyr4t1h78kquhi9bgjsi/Multimodal_all__presentation_.pdf?rlkey=pgwiufej87l09r1s283gtjbmp&st=bfhk1zp7&dl=0`.

FIGURE 1. Summary of results.

		
No advice No extra information	Weak advice Warm start for each mode	Strong advice Samples from distribution
Tempering algorithms (ST, SMC,...)	ALPS with mode rebalancing	MC with data initialization gives fresh samples
Strong assumptions (necessary)	General assumptions	Very general assumptions
[GLR18], [LS24]	[LS25]	[KLV25]

We formalize the problem as follows: Sample from $\pi(x) \propto e^{-V(x)}$ (w.r.t. a reference measure on Ω) within ε distance in total variation (TV), given query access to V (and perhaps ∇V, for $\Omega = \mathbb{R}^d$). Assume $\pi = \sum_{i=1}^m w_i \pi_i$, where each component π_i satisfies a functional inequality (Poincaré or log-Sobolev); that is, the natural Markov chain on the space (e.g. Langevin diffusion or Glauber dynamics) mixes rapidly.

1. No advice

Without extra information, guarantees are available only under strong conditions. Early work gives guarantees for simulated and parallel tempering assuming suitable decompositions [MR02; WSH09]. To state our result, we assume we are given access to a sequence of distributions π_ℓ satisfying the following.

(1) (Decomposition at each temperature) For each $\ell \in [L]$, $\pi_\ell(x) = \sum_{i=1}^m w_i^{(\ell)} \pi_i^{(\ell)}$.
(2) (Mixing of each component) Each component satisfies a Poincaré inequality with constant C_{PI} or log-Sobolev inequality with constant C_{LSI}.
(3) (Mixing at highest temperature) π_1 satisfies a Poincaré inequality with constant C_{PI} or log-Sobolev inequality with constant C_{LSI}.
(4) (Closeness between temperatures) $\chi^2(\pi_i^{(\ell+1)} \| \pi_i^{(\ell)}) = O(1)$.
(5) (Bottleneck) $\min_{\ell<\ell'} \frac{w_i^{(\ell)}}{w_i^{(\ell')}} \geq \gamma$.

A common choice for the interpolating distributions is $\pi_\ell = \pi^{\beta_\ell}$. Simulated tempering runs a Markov chain on the state space $\Omega \times [L]$.

Theorem 1.1 ([GLR18]). *Under the above conditions, the simulated tempering Markov process mixes in time polynomial in all parameters. In particular, this gives a poly-time algorithm for sampling from π on $\mathbb{R}^d$ that is a mixture of strongly log-concave distributions, $\pi_i = e^{-f_0(x-\mu_i)}$, where f_0 is strongly log-concave and smooth.*

The main proof technique is Markov chain decomposition (two-scale functional inequalities). See [GBZ25] for further results.

Another classical algorithm is Sequential Monte Carlo (SMC), which has the advantage of only moving particles through distributions uni-directionally, but is more challenging to analyze. [PJT18; MS24] show guarantees for multimodal distributions but require separation between modes. We give guarantees under the above general conditions, with two stronger conditions.

(1') (Decomposition at each temperature) For each ℓ, $\pi_\ell(x) = \sum_{i=1}^m w_i \pi_i^{(\ell)}$.
(5') (Lower bound on minimum weight) $w_{\min} = \min w_i$.

Theorem 1.2 ([LS24]). *Under the strengthened conditions, with $N = \Omega\left(L \max\left\{\frac{1}{\varepsilon^2}, \frac{1}{w_{\min}^{1/4}}\right\}\right)$ particles, running SMC for appropriate poly-time, the distribution of a sample is $\leq \varepsilon$ in TV distance from π.*

Two main ingredients in the proof are showing intra-mode variance decay and hypercontractivity for mixtures. The requirement of unchanging component weights is relaxed by [HIS25].

An inherent challenge that leads to restrictive assumptions in the above results is the following: in general, a component can have smaller weights at higher temperatures,

creating a "bottleneck" that prevents samples from moving into that mode. In simple terms, it is generally difficult to find a mode. Formally, considering a family of perturbations of two Gaussians with different covariances, no algorithm can generate a sample within constant TV distance with sub-exponentially many queries to π or $\nabla \ln \pi$ [GLR18].

2. Weak advice

As mode location is an inherent challenge, a natural assumption to isolate the search problem from the sampling problem is to assume we already have *warm starts* $\{x_j\}$ to the modes, e.g. obtained by multiple runs of optimization. [TMR21] introduce the annealed-leap point sampler (ALPS), which combines tempering towards a mixture of peaked distributions, with teleportation, and gives asymptotic analysis in the limit as the modes become gaussian [RRT22]. Using a warm start assumption, we can do away with the bottleneck assumption and give a general result. The algorithm requires choosing a tilting function q_β, for example gaussian $e^{-\frac{\beta}{2}\|x\|^2}$.

(5") (Warm start: Tilt towards x_{j_i} puts at least a constant amount of mass on the ith mode.) For each $i \in [m]$, there exists j_i such that for every $\beta > 0$,

$$\int_X \alpha_i \pi_i(x) q_\beta(x - x_{j_i}) dx \geq c_0 \int_X \pi(x) q_\beta(x - x_{j_i}) dx.$$

Theorem 2.1 ([LS25])**.** *Under these (and additional technical) assumptions, ALPS with mode rebalancing approximately samples in poly-time.*

The main algorithmic and proof challenge is estimating partition functions of components to rebalance weights between modes.

3. Strong advice

Consider strong advice in the form of a few samples from the target distribution. Although strong, this is present in the setting of generative modeling, when a dataset of samples is given and the task is to learn to generate new samples. [KLV25] show that the problem is generically solvable: for a mixture with m components, given $\tilde{O}(m/\varepsilon^2)$ samples, a fresh sample within distance ε in TV can be generated by simply running the Markov chain starting from a random sample; this is termed *data-based initialization*. In fact, the theorem works under the higher-order spectral gap assumption $\lambda_{m+1} \geq \alpha$ which is implied by being a mixture of distributions satisfying Poincaré.

Theorem 3.1 ([KLV25])**.** *Suppose $\lambda_{m+1} \geq \alpha$ and there are constants t_0, R such that*

$$(\textit{warm start after time } t_0) \quad \forall y \in \Omega, \quad \chi^2(\delta_y P_{t_0} \| \pi) \leq R.$$

Then w.p. $\geq 1 - \delta$, for $n = \Omega\left(\frac{m}{\varepsilon_{TV}^2} \ln \frac{m}{\delta}\right)$, $t \geq t_0 + \frac{1}{\alpha} \ln \frac{4R}{\varepsilon_{TV}^2}$, we have $TV(\mu_t, \pi) \leq \varepsilon_{TV}^2$.

The key notion for the proof is that of a (m, ε)-eigenfunction balanced initialization: a random sample of size $\widetilde{O}(m/\varepsilon^2)$ satisfies this with high probability by concentration bounds, and the theorem follows from this initialization by eigenfunction expansion. The theorem is robust to error in the Markov chain (so that it applies to score matching and pseudo-likelihood estimation), and gives applications to learning approximately low-rank Ising models.

This improves prior work by [KV23]; see also follow-up work by [Gay+25].

REFERENCES

[Gay+25] William Gay, William He, Nicholas Kocurek, and Ryan O'Donnell. "Sampling and Identity-Testing Without Approximate Tensorization of Entropy". In: *arXiv preprint arXiv:2506.23456* (2025).

[GBZ25] Jhanvi Garg, Krishna Balasubramanian, and Quan Zhou. "Restricted Spectral Gap Decomposition for Simulated Tempering Targeting Mixture Distributions". In: *arXiv preprint arXiv:2505.15059* (2025).

[GLR18] Rong Ge, Holden Lee, and Andrej Risteski. "Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo". In: *Advances in neural information processing systems* 31 (2018).

[HIS25] Ruiyu Han, Gautam Iyer, and Dejan Slepčev. "Polynomial complexity sampling from multimodal distributions using Sequential Monte Carlo". In: *arXiv preprint arXiv:2508.02763* (2025).

[KLV25] Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. "Efficiently learning and sampling multimodal distributions with data-based initialization". In: *Proceedings of Thirty Eighth Conference on Learning Theory*. Ed. by Nika Haghtalab and Ankur Moitra. Vol. 291. Proceedings of Machine Learning Research. PMLR, 30 Jun–04 Jul 2025, pp. 3264–3326. URL: https://proceedings.mlr.press/v291/koehler25a.html.

[KV23] Frederic Koehler and Thuy-Duong Vuong. "Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization". In: *arXiv preprint arXiv:2310.01762* (2023).

[LS24] Holden Lee and Matheau Santana-Gijzen. *Convergence Bounds for Sequential Monte Carlo on Multimodal Distributions using Soft Decomposition*. 2024. arXiv: 2405.19553 [math.ST].

[LS25] Holden Lee and Matheau Santana-Gijzen. "Sampling multimodal distributions with warm starts". In: *Submission* (2025).

[MR02] Neal Madras and Dana Randall. "Markov chain decomposition for convergence rate analysis". In: *Annals of Applied Probability* (2002), pp. 581–606.

[MS24] Joseph Mathews and Scott C Schmidler. "Finite sample complexity of sequential Monte Carlo estimators on multimodal target distributions". In: *The Annals of Applied Probability* 34.1B (2024), pp. 1199–1223.

[PJT18] Daniel Paulin, Ajay Jasra, and Alexandre H. Thiery. "Error Bounds for Sequential Monte Carlo Samplers for Multimodal Distributions". In: *The Annals of Applied Probability* 28.3 (2018), pp. 1495–1535. DOI: 10.1214/17-AAP1323.

[RRT22] Gareth O. Roberts, Jeffrey S. Rosenthal, and Nicholas G. Tawn. "Skew brownian motion and complexity of the alps algorithm". In: *Journal of Applied Probability* 59.3 (2022), pp. 777–796. DOI: 10.1017/jpr.2021.78.

[TMR21] Nicholas G Tawn, Matthew T Moores, and Gareth O Roberts. "Annealed Leap-Point Sampler for multimodal target distributions". In: *arXiv preprint arXiv:2112.12908* (2021).

[WSH09] Dawn Woodard, Scott Schmidler, and Mark Huber. "Conditions for Rapid Mixing of Parallel and Simulated Tempering on Multimodal Distributions". In: *The Annals of Applied Probability* 19 (June 2009). DOI: 10.1214/08-AAP555.

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS, 3400 NORTH CHARLES STREET, BALTIMORE, MD 21218

Email address: hlee283@jhu.edu

INVERSE PROBLEM OVER PROBABILITY MEASURE SPACE

QIN LI

Classification AMS 2020: 35R30, 65Jxx, 60B05, 90Cxx

Keywords: inverse problems, probability measures, optimization

Inverse problems are ubiquitous. Traditionally, the goal is to infer an unknown vector or function. We study if we can study inverse problems over probability measure space.

To be more specific, a classical inverse problem can be posed as:

$$\text{find} \quad x \quad \text{so that} \quad \mathcal{G}(x) \approx y$$

where y is the data and $\mathcal{G}$ is the forward map. Here x and y can either be objects living in a finite dimensional space, such as vectors in $\mathbb{R}^d$, or infinite dimensional space such as $L_2(\Omega)$ function space. We examine the problem that levies the question up to the probability measure space:

$$\text{find} \quad \rho_x \quad \text{so that} \quad \mathcal{G}_{\#}\rho_x \approx \rho_y$$

where ρ_y is a probability measure of data. It can either be a point cloud so that $\rho_y = \frac{1}{N}\sum_i \delta_{y_i}$ or something with a smooth density $\rho_y = p_y dy$.

The problem naturally arises in many continuous interpretation of machine learning algorithms where a probability measure is to be reconstructed. Mean-field limit of neural network training, for example, can be framed as an inverse problem over the probability space.

Depending on the over- and under-determinedness of the system, the formulation is accordingly adjusted. In the overdetermined case, we look for the optimizer through

$$\min_{\rho_x} \mathcal{D}\left(\mathcal{G}_{\#}\rho_x \,, \rho_y\right) \,.$$

It turns out

- Setting $\mathcal{D}$ to be any ϕ divergence, the reconstruction is the conditional distribution;
- Setting $\mathcal{D}$ to be any Wasserstein distance, the reconstruction is the marginal distribution.

In the underdetermined case, we look for the optimizer through

$$\min_{\mathcal{G}_{\#}\rho_x = \rho_y} \mathcal{E}\left(\rho_x\right) \,.$$

It turns out

- Setting $\mathcal{E}$ to be entropy, the solution is piecewise constant;
- Setting $\mathcal{E}$ to be of moments, the optimizer is generated by least-norm solution.

These finding suggests that Wasserstein is a very close counterpart of Euclidean norm, while entropy introduces very different structural reconstruction [2].

We finally discuss some optimization solvers that finds these solution. In particular, beyond the gradient flow, we also discussed the Hamiltonian flow, which in theory should give faster convergence rate in the continuous-in-time setting [1].

REFERENCES

[1] Shi Chen, Qin Li, Oliver Tse and Stephen Wright Accelerating optimization over the space of probability measures *Journal of Machine Learning Research*, 26, 1-40, 2025

[2] Qin Li, Maria Oprea, Yunan Yang and Li Wang Inverse Problems Over Probability Measure Space *arXiv*, 2504.18999

480 LINCOLN DR., MADISON, WI

Email address: qinli@math.wisc.edu

PROPAGATION OF CHAOS FOR MEAN-FIELD LANGEVIN DYNAMICS

ATSUSHI NITANDA

Classification AMS 2020: 60K35, 65K10, 90C26

Keywords: mean-field Langevin dynamics, neural network, interacting particle systems

A two layer *mean-field neural network* (MFNN) with N neurons is defined as an empirical average of N functions: $\mathbb{E}_{X\sim\rho_{\mathbf{x}}}\left[h\left(X,\cdot\right)\right]=\frac{1}{N}\sum_{i=1}^{N}h(x^i,\cdot)$, where each $h(x^i,\cdot)$ represents a single neuron with parameter x^i and $\rho_{\mathbf{x}}=\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ is an empirical distribution. As the number of neurons get infinitely large $(N\to\infty)$, the *mean-field limit* is attained: $\rho_{\mathbf{x}}\to\mu$, leading to MFNN having an infinite number of particles: $\mathbb{E}_{X\sim\mu}\left[h(X,\cdot)\right]$. Since a distribution μ parameterizes the model in this mean-field limit, training can now be formulated as the optimization over the space of probability distributions [Nitanda and Suzuki, 2017]. Gradient descent for MFNNs exhibits global convergence [Chizat and Bach, 2018, Mei et al., 2018] and adaptivity [Yang and Hu, 2020, Ba et al., 2022]. To improve stability during training, one may consider *noisy* gradient training by adding Gaussian noise, giving rise to *mean-field Langevin dynamics* (MFLD) [Mei et al., 2018, Hu et al., 2019]. MFLD, with $N=\infty$, also achieves global convergence to the optimal solution [Hu et al., 2019, Jabir et al., 2019], with an exponential convergence rate under the *uniform log-Sobolev inequality* (LSI) Nitanda et al. [2022], Chizat [2022] in the continuous-time setting.

However, the mean-field limit attained at $N=\infty$ cannot be accurately replicated in real-life scenarios. When employing a finite-particle system $\rho_{\mathbf{x}}$, the approximation error that arises has been studied in the literature on *propagation of chaos* (PoC) Sznitman [1991]. In the context of MFLD, Chen et al. [2022], Suzuki et al. [2023] proved the *uniform-in-time* PoC for the trajectory of MFLD. In particular, in the long-time limit, they established the bounds $\mathcal{L}^{(N)}(\mu_*^{(N)})-\mathcal{L}(\mu_*)=O\left(\frac{\lambda}{\alpha N}\right)$, where $\alpha\gtrsim\exp\left(-\Theta\left(\frac{1}{\lambda}\right)\right)$ is the LSI constant on *proximal Gibbs distributions*, λ is the regularization coefficient, and $\mathcal{L}^{(N)}(\mu_*^{(N)})$ and $\mathcal{L}(\mu_*)$ are the optimal values in finite- and infinite-particle systems. Subsequently, Nitanda [2024] improved upon this result by removing α from the above bound, resulting in $O\left(\frac{1}{N}\right)$. This refinement of the bound is significant as previously, the LSI constant could become exponentially small as $\lambda\to0$. While Nitanda [2024] also established PoC for the MFLD trajectory by incorporating the *uniform-in-N* LSI Chewi et al. [2024]: $\mathcal{L}^{(N)}(\mu_t^{(N)})\to\mathcal{L}^{(N)}(\mu_*^{(N)})$, this approach is indirect for showing convergence to the mean-field limit $\mathcal{L}(\mu_*)$ and results in a slower convergence rate over time.

In this work, we further aim to improve PoC for MFLD by demonstrating a faster convergence rate in time, while maintaining the final approximation error $O\left(\frac{1}{N}\right)$ attained at $t=\infty$. We then utilize our result to propose a PoC-based ensemble technique by demonstrating how finite particle systems can converge towards the mean-field limit when merging MFNNs trained in parallel.

0.1. **Contributions.** The PoC for MFLD Chen et al. [2022], Suzuki et al. [2023] consists of particle approximation error $O\left(\frac{\lambda}{\alpha N}\right)$ due to finite-N-particles and optimization error $\exp(-\Theta(\lambda\alpha t))$. This result basically builds upon the defective LSI: $\exists\delta > 0$,

$$\frac{1}{N}\mathcal{L}^{(N)}(\mu^{(N)}) - \mathcal{L}(\mu_*) \leq \frac{\delta}{N} + \frac{\lambda}{2\alpha N}\mathrm{FI}(\mu^{(N)}\|\mu_*^{(N)})$$

implicitly established by Chen et al. [2022] under the uniform LSI condition Nitanda et al. [2022], Chizat [2022], where FI is Fisher information. The dependence on LSI-constant α in $O\left(\frac{\lambda}{\alpha N}\right)$ of PoC is basically inherited from δ. In our work, we first remove the dependence on α from δ by introducing *uniform directional LSI* in training MFNNs setting. Based on the improved defective LSI, we then derive an improved PoC for MFLD where the particle approximation error is $O\left(\frac{1}{N}\right)$ as follows.

Theorem 0.1 (Propagation chaos for MFLD)**.** *Under the unifrom directional LSI with a constant $\alpha > 0$ and regular conditions. Then, MFLD in the continuous-time satisfies*

$$\frac{1}{N}\mathcal{L}^{(N)}(\mu_t^{(N)}) - \mathcal{L}(\mu_*) \leq \frac{B}{N} + \exp(-2\alpha\lambda t)\Delta_0^{(N)}.$$

Similar to Nitanda [2024], this improvement exponentially reduces the required number of particles since the constant $\alpha \gtrsim \exp\left(-\Theta(\frac{1}{\lambda})\right)$ can exponentially decrease as $\lambda \to \infty$. Moreover, our result demonstrates a faster optimization speed compared to Nitanda [2024] due to a different exponent α in the optimization error terms: $\exp(-\Theta(\lambda\alpha t))$. In our analysis, α is a constant of the uniform directional LSI, which is larger than the LSI constant on $\mu_*^{(N)}$ appearing in the optimization error in Nitanda [2024].

References

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems 35*, 2022.

Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

Sinho Chewi, Atsushi Nitanda, and Matthew S Zhang. Uniform-in-n log-sobolev inequality for the mean-field langevin dynamics with convex energy. *arXiv preprint arXiv:2409.10440*, 2024.

Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3040–3050, 2018.

Kaitong Hu, Zhenjie Ren, David Siska, and Lukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

Jean-François Jabir, David Šiška, and Łukasz Szpruch. Mean-field neural odes via relaxed optimal control. *arXiv preprint arXiv:1912.05475*, 2019.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Atsushi Nitanda. Improved particle approximation error for mean field neural networks. In *Advances in Neural Information Processing Systems 37*, 2024.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *Proceedings of International Conference on Artificial Intelligence and Statistics 25*, pages 9741–9757, 2022.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. In *Advances in Neural Information Processing Systems 36*, 2023.

Alain-Sol Sznitman. Topics in propagation of chaos. *Ecole d'Eté de Probabilités de Saint-Flour XIX—1989*, pages 165–251, 1991.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Centre for Frontier AI and Research, A⋆STAR, College of Computing and Data Science, Nanyang Technological University

Email address: atsushi_nitanda@a-star.edu.sg

WASSERSTEIN FISHER RAO GRADIENT FLOWS: SMC & OPERATOR SPLITTING

SAHANI PATHIRAJA

Classification AMS 2020: 65B99, 65C05, 82M31, 60G35, 62C10

Keywords: Gradient Flows, Partial Differential Equations, Monte Carlo algorithms, Stochastic Filtering

This talk discusses three related pieces of work, primarily built on connections between stochastic filtering, sampling and evolutionary dynamics. The first part discusses recent work [1] builds on a long-standing view that there is a connection between the dynamical equations describing evolutionary processes in biology and sequential Bayesian learning methods. The paper [1] describes new research in which this precise connection is rigorously established in the continuous time setting, where previously this was done in discrete time. We presented a detailed investigation of connections between continuous time, continuous trait Crow-Kimura replicator-mutator dynamics [2] and the fundamental equation of non-linear filtering, the KushnerStratonovich partial differential equation (PDE). Inspired by a non-local fitness functional presented in the mathematical biology literature [3], we extended this connection to obtain a "modified" Kushner-Stratonovich equation. This equation was shown to beneficial for filtering with misspecified models and a specific choice of parameters in the fitness functional was shown to coincide with covariance inflated Kalman Bucy filtering, in the linear-Gaussian setting. Additionally, we considered the misspecified model filtering problem, with linear-Gaussian dynamics and where the misspecification arises through an unknown constant bias in the signal dynamics. We proved that through a judicious choice of parameters in the fitness functional, mean squared error and uncertainty quantification (through the covariance) could be improved via this modified Kushner-Stratonovich equation. Estimation is improved over traditional covariance inflation techniques, as well as over the standard filtering setup (assuming perfect model knowledge). There are several avenues for further work, most notably, the analysis on misspecified models which has primarily focused on the scalar setting which has simplified the analysis. In future works, the multivariate setting, as well as extensions to nonlinear dynamics should be explored.

The second part discusses recent work [4] that focuses on the related problem of sampling from an unnormalised target distribution of the form $\pi \propto e^{-V}$. It is well known that this sampling problem can be written as an optimisation problem over the space of probability distribution in which we aim to minimise the Kullback–Leibler divergence to π. Doing so allows to derive partial differential equations that are gradient flows of the Kullback-Leibler divergence to π, which can be formulatd using either the Wasserstein, Fisher-Rao or Wasserstein-Fisher-Rao metrics. The latter in particular can be interpreted as a replicator-mutator equation, with the main difference to the first part being that the fitness function is now static, but dependent on the

solution of the PDE ρ_t. We additionally connected these gradient flows to several known sequential Monte Carlo algorithms in the literature.. We focused in particular on PDEs obtained by considering the Wasserstein-Fisher-Rao geometry over the space of probabilities and show that these lead to a natural implementation using importance sampling and sequential Monte Carlo. We proposed a novel algorithm to approximate the Wasserstein–Fisher–Rao flow of the Kullback–Leibler divergence which empirically outperforms the current state-of-the-art.

The third part discussed ongoing work on operator splitting of the Wasserstein-Fisher-Rao gradient flow PDE. As this PDE consists of the sum of the Wasserstein and Fisher-Rao operators, a natural approach to numerically solving this PDE is via operator splitting. We demonstrated that the order of solving the two operators (Wasserstein first, then Fisher-Rao vs Fisher-Rao first then Wasserstein) induces biases which can be exploited to improve the speed of convergence to π, even compared to the exact solution of the Wasserstein-Fisher-Rao gradient flow PDE. Such biases do not affect the invariant density, only the path of intermediate densities taken to reach the invariant density. We quantified this behaviour in the Gaussian case, and showed that specific pairings of initial and target densities require a specific ordering of operators to achieve a speed-up. Some open problems related to proving this phenomenon in the more general setting were discussed.

REFERENCES

[1] Pathiraja, Sahani & Wacker, Philipp. Connections between sequential Bayesian inference and evolutionary dynamics *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* , Volume 383, Issue 2298, 2025.

[2] Kimura, M. A stochastic model concerning the maintenance of genetic variability in quantitative characters *Proceedings of the National Academy of Sciences*, 54, 731-736, 1965

[3] Cressman R, Hofbauer J, Riedel F . Stability of the replicator equation for a single species with a multi-dimensional continuous trait space. J *Journal of Theoretical Biology*, 239, 273-288, 2006

[4] Crucinio,Francesca & Pathiraja, Sahani. Sequential Monte Carlo approximations of Wasserstein–Fisher–Rao gradient flows *arXiv*, https://arxiv.org/pdf/2506.05905, 2025.

SCHOOL OF MATHEMATICS AND STATISTICS, UNSW SYDNEY, AUSTRALIA
Email address: s.pathiraja@unsw.edu.au

MCKEAN–PONTRYAGIN MINIMUM PRINCIPLE FOR STOCHASTIC OPTIMAL CONTROL

SEBASTIAN REICH

Classification AMS 2020: 35F21, 49M99, 93E20, 70H30, 70H45

Keywords: Stochastic optimal control, Pontryagin minimum principle, McKean evolution equations

Summary. In this talk, a deterministic mean-field formulation of the Pontryagin minimum principle for stochastic optimal control problems has been sketched out following our recent technical report [3]. Contrary to the well-known forward and backward SDE formulation of the stochastic Pontryagin minimum principle [1], the proposed mean-field approach leads to a gauge variable which can be freely chosen and can be used to decouple the arising forward and reverse time mean-field ODEs.

Problem statement. We consider the optimal control problem for a controlled SDE of the form

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + GU_t\mathrm{d}t + \Sigma^{1/2}\mathrm{d}B_t, \qquad X_0 = a, \tag{0.1}$$

under finite horizon cost function

$$J_T(a, U_{0:T}) = \mathbb{E}\left[\int_0^T \left(c(X_t) + \frac{1}{2}U_t^{\mathrm{T}} R^{-1} U_t\right)\mathrm{d}t + f(X_T)\right]. \tag{0.2}$$

Here B_t denotes d_x-dimensional Brownian motion, $\Sigma \in \mathbb{R}^{d_x \times d_x}$ the symmetric positive definite diffusion matrix, $R \in \mathbb{R}^{d_u \times d_u}$ a symmetric positive definite weight matrix, $G \in \mathbb{R}^{d_x \times d_u}$ the control matrix, $c(x)$ the running cost, and $f(x)$ the terminal cost. See, for example, reference [1] for more details. We also introduce the weighted norm $\|\cdot\|_R$ via $\|u\|_R^2 = u^{\mathrm{T}} R^{-1} u$.

The aim is to find the closed loop control law $U_t = u_t(X_t)$ that minimizes $J_T(a, U_{0:T})$ over the set of admissible control laws. It is well-known [1] that, assuming sufficient regularity, the desired closed loop control law is provided by

$$u_t(x) = -RG^{\mathrm{T}}\nabla_x v_t(x) \tag{0.3}$$

with the optimal value function $v_t(x)$ satisfying the Hamilton–Jacobi–Bellman (HJB) equation

$$-\partial_t v_t = b \cdot \nabla_x v_t + \frac{1}{2}\Sigma : D_x^2 v_t + c + \min_u \left(Gu \cdot \nabla_x v_t + \frac{1}{2}\|u\|_R^2\right), \quad v_T = f. \tag{0.4}$$

McKean–Pontryagin minimum principle. We now formulate the proposed McKean–Pontryagin minimum principle. The initial conditions $a \in \mathbb{R}^{d_x}$ may be viewed as a label in the sense of Lagrangian fluid dynamics, which we assume to be distributed according

to a probability density function π_0. We therefore consider functions $x(a)$, $p(x)$, $u(a)$, and $\beta(a)$ and introduce the Hamiltonian functional

$$\mathcal{H}(x,p,u,\beta) = \int_{\mathbb{R}^{d_x}} H(x,p,u,\beta)(a)\,\pi_0(a)\,\mathrm{d}a \tag{0.5}$$

with Hamiltonian density

$$H(x,p,u,\beta)(a) := p(a)^{\mathrm{T}}\left(b(x(a)) + Gu(a)\right) + \frac{1}{2}\nabla_x \cdot (\Sigma\phi(x(a))) + \tag{0.6a}$$

$$\beta(a)^{\mathrm{T}}\left(p - \phi(x(a))\right) + c(x(a)) + \frac{1}{2}\|u(a)\|_R^2. \tag{0.6b}$$

Here $\beta(a) \in \mathbb{R}^{d_x}$ takes the role of a gauge variable [2], which does not appear in the classical Pontryagin minimum principle [4]. We also note the occurrence of the function $\phi(x)$, which will be determined in terms of the non-holonomic constraint arising from variations with respect to $\beta(a)$ [2]. More specifically, the desired equations of motion are induced by the phase space action principle [2] applied to

$$\mathcal{S} = \int_{\mathbb{R}^{d_x}} \left\{ \int_0^T \left(P_t^{\mathrm{T}} \dot{X}_t - H(X_t, P_t, U_t, \beta_t) \right) \mathrm{d}t - f(X_T) \right\} \pi_0 \mathrm{d}a. \tag{0.7}$$

Taking variations with respect to U_t, we find that the optimal control satisfies

$$\nabla_u H(X_t(a), P_t(a), U_t(a), \beta_t(a)) = R^{-1}U_t(a) + G^{\mathrm{T}}P_t(a) = 0. \tag{0.8}$$

Variations with respect to β_t lead on the other hand to the constraint

$$P_t(a) - \phi_t(X_t(a)) = 0, \tag{0.9}$$

which defines the function $\phi_t(x)$ in terms of $X_t(a)$ and $P_t(a)$. Using the thus specified $\phi_t(x)$, we obtain the closed loop control

$$u_t(x) = -RG^{\mathrm{T}}\phi_t(x). \tag{0.10}$$

Finally, variations with respect to X_t and P_t lead to the Hamiltonian evolution equations in (X_t, P_t); i.e.,

$$\dot{X}_t(a) = +\nabla_p H(X_t(a), P_t(a), U_t(a), \beta_t(a)), \tag{0.11a}$$

$$\dot{P}_t(a) = -\nabla_x H(X_t(a), P_t(a), U_t(a), \beta_t(a)) \tag{0.11b}$$

for each $a \in \mathbb{R}^{d_x}$. The boundary conditions are $X_0(a) = a \sim \pi_0$ and $P_T(a) = \nabla_x f(X_T(a))$. Dropping the label $a \in \mathbb{R}^{d_x}$ from now on, the Hamiltonian equations of motion (0.11) therefore become

$$\dot{X}_t = b(X_t) + GU_t + \beta_t, \tag{0.12a}$$

$$\dot{P}_t = \left(D_x\phi_t(X_t)\right)^{\mathrm{T}}\beta_t - \left(D_x b(X_t)\right)^{\mathrm{T}} P_t - \frac{1}{2}\nabla_x\nabla_x \cdot (\Sigma\phi_t(X_t)) - \nabla_x c(X_t). \tag{0.12b}$$

The following theorem provides the key result with regard to the gauge variable β_t and demonstrates that (0.12) indeed delivers the desired extension of the classical Pontryagin minimum principle to stochastic optimal control problems.

Theorem 0.1. *For any choice of the gauge variable β_t, the resulting function $\phi_t(x)$ satisfies*

$$\phi_t(x) = \nabla_x v_t(x) \tag{0.13}$$

where $v_t(x)$ is the value function satisfying the HJB equation (0.4).

Proof. Let us derive the evolution equation for $\phi_t(x)$ implied by (0.9):

$$-\partial_t \phi_t(X_t) = D_x\phi_t(X_t)\dot{X}_t - \dot{P}_t \tag{0.14a}$$

$$= D_x\phi_t(X_t)\left(b(X_t) + GU_t\right) + \frac{1}{2}\nabla_x\nabla_x\cdot(\Sigma\phi_t(X_t)) + \tag{0.14b}$$

$$\left(D_x b(X_t)\right)^{\mathrm{T}}\phi_t(X_t) + \nabla_x c(X_t). \tag{0.14c}$$

Here we have used that $D_x\phi_t(x)$ is symmetric since $\phi_t(x)$ itself is the gradient of the value function $v_t(x)$. Hence, $\phi_t(x)$ satisfies the reverse time PDE

$$-\partial_t\phi_t = D_x\phi_t\left(b + GU_t\right) + \frac{1}{2}\nabla_x\nabla_x\cdot(\Sigma\phi_t) + \left(D_x b\right)^{\mathrm{T}}\phi_t + \nabla_x c \tag{0.15}$$

subject to the terminal condition $\phi_T = \nabla_x f$, which also follows from (0.4) by taking the gradient. Hence $\phi_t(x) = \nabla_x v_t(x)$ independent of β_t. □

A natural choice for the gauge function β_t is

$$\beta_t = -\frac{1}{2}\Sigma\nabla_x \log\pi_t(X_t), \tag{0.16}$$

where $\pi_t(x)$ denotes the law of X_t. Alternatively, consider

$$\beta_t = GRG^{\mathrm{T}}\phi_t(X_t), \tag{0.17}$$

which eliminates the control from the forward evolution equation in X_t since

$$GU_t = -GRG^{\mathrm{T}}\phi_t(X_t) = -\beta_t. \tag{0.18}$$

Both choices for the gauge variable can also be combined into

$$\beta_t = GRG^{\mathrm{T}}\phi_t(X_t) + Gu_t^{\mathrm{ref}}(X_t) - \frac{1}{2}\Sigma\nabla_x\log\pi_t(X_t), \tag{0.19}$$

where $u_t^{\mathrm{ref}}(x)$ denotes a known reference control; if available.

Applications. Numerical results and an extension to infinite horizon discounted cost functionals can be found in the report [3]. An application of the proposed methodology to model predictive control [5] can also be found in [3].

References

[1] René Carmona. *Lectures on BSDEs, Stochastic Control, and Stochastic Differential Games with Financial Applications*. SIAM, Philadelphia, 2016.

[2] Paul A.M. Dirac. Generalized Hamiltonian dynamics. *Can. J. Math.*, 2:129–148, 1950.

[3] M. Opper and S. Reich. On a mean-field Pontryagin minimum principle for stochastic optimal control. *arXiv:2506.10506*, 2025.

[4] L.S. Pointryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mihchenko. *The Mathematical Theory of Optimal Processes*. John Wiley & Sons, New York, 1962.

[5] James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, Madison, 2nd edition, 2018.

Department of Mathematics, University of Potsdam, D-14476 Potsdam, Germany
Email address: sebastian.reich@uni-potsdam.de

UNIFORM ERROR BOUNDS OF THE ENSEMBLE SQUARE ROOT FILTER FOR CHAOTIC DYNAMICS WITH MULTIPLICATIVE COVARIANCE INFLATION

KOTA TAKEDA

Classification AMS 2020: 65C05, 35R30, 35Q93, 62M20, 62F15

Keywords: data assimilation, dissipative dynamics, ensemble Kalman filter, accuracy, filtering problem

Data assimilation aims to estimate the hidden true state from noisy observations by utilizing background model dynamics. As a model dynamics, we consider a class of nonlinear dynamical systems on Hilbert spaces including the two-dimensional Navier-Stokes equations and the Lorenz 63 and 96 models. For nonlinear model dynamics, the ensemble Kalman filter (EnKF) is often used to approximate the mean and covariance of the probability distribution with a set of particles called an ensemble. There are two major variants of the EnKF: a stochastic one is the Perturbed Observation (PO) method and a deterministic one is the Ensemble Transform Kalman Filter (ETKF). The PO method is simple to implement but suffers from sampling errors due to the perturbation of observations. On the other hand, the ETKF avoids such sampling errors and often outperforms the PO method [1]. Recent theoretical studies reveal the basic properties of the EnKF [2, 3, 4]. While these basic results have been established in general settings, the long-time accuracy of the EnKF is studied in limited situations. The uniform-in-time error bound for the PO method has been obtained under suitable conditions [5]. However, such a bound for the ETKF has not been established yet due to difficulty in analyzing the complicated update step of the ETKF.

In this talk, we show that the uniform-in-time error bound for the ETKF is obtained when the system is finite-dimensional [6], i.e., the state estimation error $\boldsymbol{\delta}_n$ at a time step $n \in \mathbb{N}$ of the ETKF satisfies

$$\limsup_{n\to\infty} \mathbb{E}[|\boldsymbol{\delta}_n|^2] \leq C\gamma^2,$$

where $\gamma > 0$ is the standard deviation of random observation noises and $C > 0$ is a constant independent of n and γ. The other conditions are explained in the talk and the full paper [6]. This bound justifies that the ETKF can accurately estimate the true state from noisy observations over long time intervals when the observation noise is sufficiently small. We also show numerical experiments with the Lorenz 96 model to demonstrate the validity of our result.

REFERENCES

[1] Andrew J. Majda, John Harlim Filtering complex turbulent systems. *Cambridge University Press*, 2012.

[2] Evan Kwiatkowski, Jan Mandel Convergence of the Square Root Ensemble Kalman Filter in the Large Ensemble Limit. *SIAM/ASA Journal on Uncertainty Quantification*, 3 1, 2166–2525, 2015.

[3] Omar Al-Ghattas, Daniel Sanz-Alonso Non-asymptotic analysis of ensemble Kalman updates: effective dimension and localization. *Information and Inference: A Journal of the IMA*, 13 1, iaad043, 2015.

[4] Xin T. Tong, Andrew J. Majda, David Kelly Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity*, 29 2, 657–691, 2016.
[5] David Kelly, Kody J. H. Law, Andrew M. Stuart Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27 10, 2579–2603, 2014.
[6] Kota Takeda, Takashi Sakajo Uniform Error Bounds of the Ensemble Transform Kalman Filter for Chaotic Dynamics with Multiplicative Covariance Inflation. *SIAM/ASA Journal on Uncertainty Quantification*, 12 4, 1215–1335, 2024.

NAGOYA UNIVERSITY, FUROCHO, CHIKUSA-KU, NAGOYA, AICHI, 464-8602, JAPAN
Email address: takeda@na.nuap.nagoya-u.ac.jp

OPTIMAL RIEMANNIAN METRIC FOR POINCARÉ INEQUALITIES AND HOW TO IDEALLY PRECONDITION LANGEVIN DYNAMICS

OLIVIER ZAHM

Classification AMS 2020: 46N30, 60H10; 39B62.

Keywords: Poincaré inequality, Langevin Dynamic, Stein Kernel, moment measure.

INTRODUCTION AND PROBLEM SET-UP

We consider a generalisation of the classical Poincaré inequality to a Riemannian (or weighted-matrix) setting, and uses this to design optimal preconditioners for Langevin dynamics. Recall the usual Poincaré inequality: for a probability measure μ on $\mathbb{R}^d$, one says it satisfies a Poincaré inequality with constant C_{P} if

$$\mathrm{Var}_\mu(f) \;\leq\; C_{\mathrm{P}} \int \|\nabla f(x)\|^2 \, \mu(dx),$$

holds for all sufficiently smooth f. It is well known that this constant governs the exponential convergence rate of the overdamped Langevin SDE

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dB_t,$$

whose invariant measure is $\mu(dx) \propto e^{-U(x)}dx$. In fact, μ satisfies the above inequality with constant C_{P} if and only if the dynamics converge at rate at least $1/C_{\mathrm{P}}$.

We introduce a *Riemannian metric* $W(x)$, i.e. a symmetric positive-definite matrix-field, so that the Poincaré inequality becomes

$$\mathrm{Var}_\mu(f) \;\leq\; C_W \int \nabla f(x)^\top \, W(x) \, \nabla f(x) \; \mu(dx).$$

Here the aim is to choose $W(x)$ so as to make the constant C_W as small as possible (ideally equal to 1). Equivalently one can seek

$$\min_{W(\cdot)\succ 0} \; C_W$$

subject to a normalisation constraint

$$\int \mathrm{trace}(W)\mu(dx) = \mathrm{trace}(\mathrm{Cov}_\mu),$$

so as to avoid the trivial scaling freedom $W \mapsto \alpha W$. In doing so, one effectively finds the optimal local anisotropic diffusion for the associated Riemannian Langevin diffusion

$$dX_t = -W(X_t)\, \nabla U(X_t)\, dt + \nabla\cdot W(X_t)\, dt + \sqrt{2\, W(X_t)}\, dB_t,$$

which accelerates convergence when compared to the standar Langevin dynamic $W = I_d$.

MAIN RESULTS

Under the assumption that μ is a *moment measure*, we show that an optimal metric $W^*(x)$ exists and achieves $C_{W^*} = 1$. In particular we show that W^* can be expressed as

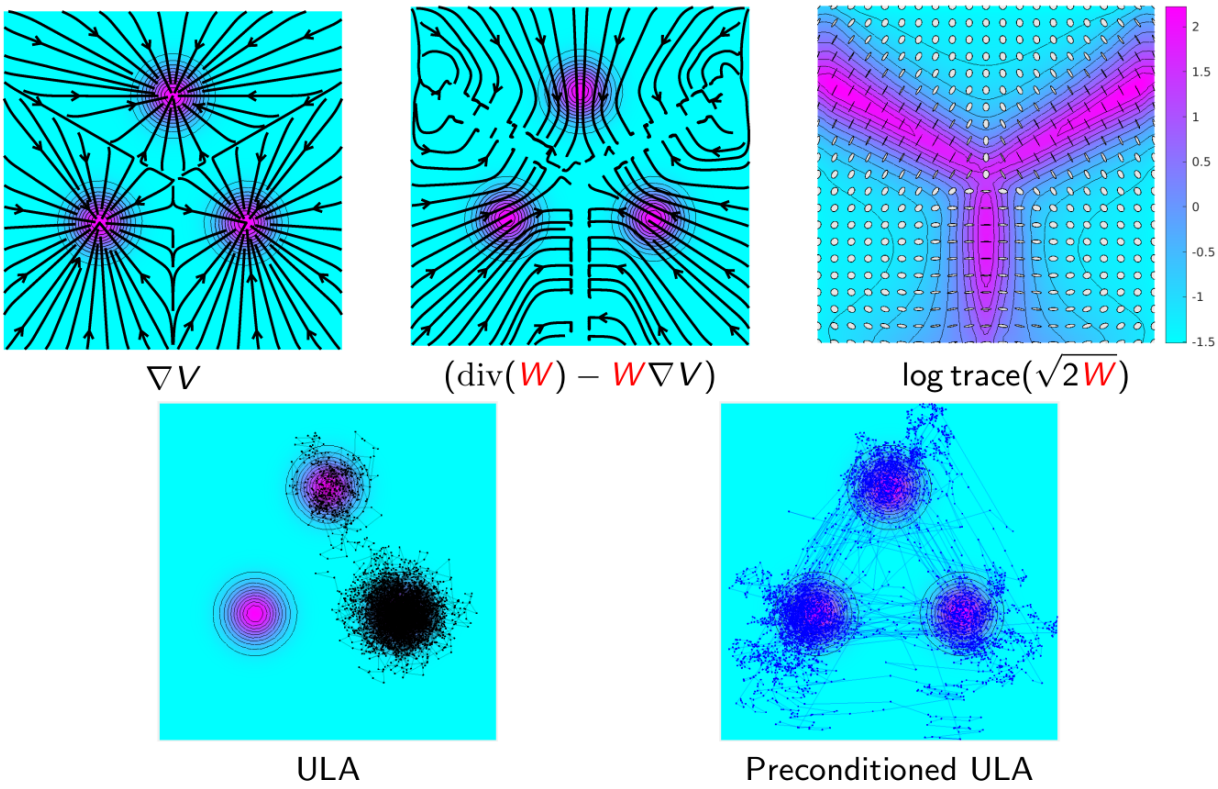
$$W^*(x) \;=\; \nabla^2\varphi(\nabla\varphi^{-1}(x))$$

for some strongly convex function φ. A key structural result is that any optimal $W^*(x)$ is a *SPD Stein kernel* of the measure μ, and so that the preconditionned Langevin dynamic simplifies as

$$dX_t = -(X_t - m)dt + \sqrt{2\,W^*(X_t)}\,dB_t,$$

where $m = \mathbb{E}_\mu[X]$ is the mean of μ.

We also show that the optimal metric can be computed numerically by solving a convex optimization problem. Using the finite element method permit to compute W^* for any measure μ in dimension $d = 2$. The next picture shows the optimal metric W^* when μ is a Gaussian mixture with 3 components.



SUMMARY OF MAIN TAKE-AWAYS

- The classical Poincaré inequality can be improved by introducing a matrix-field metric $W(x)$.
- Under moment-measure assumptions, there is an optimal metric $W^*(x)$ attaining Poincaré constant $C_{W^*} = 1$.
- This metric is essentially a Stein kernel of μ, providing a new view of Stein kernels in terms of optimal spectral gap.

- One can formulate the search for the optimal metric as a convex optimisation in the space of matrix-fields, solve it numerically by discretisation + gradient methods.
- The resulting metric can be used to precondition Langevin dynamics, giving improved convergence rates in practice.

REFERENCES

[1] Cui, Tiangang, Xin Tong, and Olivier Zahm. Optimal Riemannian metric for Poincaré inequalities and how to ideally precondition Langevin dynamics. *arXiv* 2404.02554 (2024),

INRIA GRENOBLE, FRANCE
Email address: olivier.zahm@inria.fr

DISTRIBUTION REGRESSION WITH DEEP NEURAL NETWORKS

DING-XUAN ZHOU

Classification AMS 2020: 68T07, 68Q32

Keywords: Transformer; Deep neural networks; Distribution regression; Wasserstein space of probability measures; Two-stage sampling

Deep neural networks with structures have been applied in many fields to various problems of learning from vectors. In particular, transformers have provided breakthroughs in learning tasks involving natural language processing.

In this talk, we discuss the topic of distribution regression to learn a function from the Wasserstein space $\mathcal{U} := (\mathcal{P}(\Omega), W_2)$ of probability measures on $\Omega \subset \mathbb{R}^d$ to $\mathbb{R}$. A special feature of distribution regression is two-stage sampling, meaning that the available sample for learning is not one $D = \{(\mu_i, y_i)\}_{i=1}^m$ drawn from a Borel probability measure ρ on $\mathcal{Z} = \mathcal{U} \times \mathbb{R}$, but a second-stage sample

$$\hat{D} = \left\{ \left(\{x_{i,j}\}_{j=1}^{n_i}, y_i \right) \right\}_{i=1}^m$$

with $\{x_{i,j} \in \Omega\}_{j=1}^{n_i}$ drawn from μ_i for each i. Then the empirical risk minimization over a hypothesis space $\mathcal{H}$ of continuous functions on $\mathcal{P}(\Omega)$ for learning the regression function for distribution regression $f_\rho(\mu) = \int_{\mathcal{Y}} y d\rho(y|\mu)$ defined on $\mathcal{U}$ takes the form

$$f_{\hat{D},\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \left(f(\hat{\mu}_i^{n_i}) - y_i \right)^2 .$$

A classical approach for distribution regression is kernel mean embedding with a continuous positive semi-definite (Mercer) kernel $k : \Omega \times \Omega \to \mathbb{R}$ which embed $\mu \in \mathcal{P}(\Omega)$ to a function k_μ in the reproducing kernel Hilbert space $\mathcal{H}_k$ induced by k defined as $k_\mu = \int_\Omega k(\cdot, x) d\mu(x)$. Then one can apply the kernel methods to solve the learning problem for distribution regression [1].

We apply deep neural networks to distribution regression. For an input distribution $\mu \in \mathcal{P}(\Omega)$, a deep neural network for distribution regression $\{h^{(j)} : \mathcal{P}(\Omega) \to \mathbb{R}^{d_j}\}_{j=J_1}^J$ of type (J_1, J), with depth $J \in \mathbb{N}$ and realizing level $J_1 \in \{0, 1, ..., J\}$, and width $\{d_j\}_{j=1}^J$ is defined by

$$h^{(J_1)}(\mu) = \int_\Omega \sigma \left(F^{(J_1)} \sigma \left(\cdots \sigma \left(F^{(1)} x - b^{(1)} \right) \cdots \right) - b^{(J_1)} \right) d\mu(x), \tag{0.1}$$

and $h^{(j)}(\mu) = \sigma \left(F^{(j)} h^{(j-1)}(\mu) - b^{(j)} \right), \qquad j = J_1 + 1, \ldots, J$, where $F^{(j)} \in \mathbb{R}^{d_j \times d_{j-1}}$ is a connection matrix with $d_0 = d$, and $b^{(j)} \in \mathbb{R}^{d_j}$ is a bias vector.

The case with $J = 2, J_1 = 1$ was introduced in [2].

When the target regression function takes a composite form with a polynomial Q and a univariate function g, we can take $J = 3$, $J_1 = 2$ and $\mathcal{H} = \left\{ c \cdot h^{(J)}(\mu) : \|F^{(j)}\|_\infty \le RN^2, \ \|b^{(j)}\|_\infty \le R, \ \|c\|_\infty \le RN \right\}$. We assume $|y| \le M$ almost surely and take the projection π_M onto $[-M, M]$ of the learned function $f_{\hat{D},\mathcal{H}}$.

Then the following learning rates with the error measured by the L^2 norm $\|\cdot\|_{L^2_{\rho_{\mathcal{U}}}}$ with respect to the marginal distribution $\rho_{\mathcal{U}}$ of ρ on $\mathcal{U} = \mathcal{P}(\Omega)$ were derived in [3].

Theorem 0.1. *Assume $f_\rho(\mu) = f\left(\int_\Omega g\left(Q(x)\right) d\mu(x)\right)$ for $\mu \in \mathcal{P}(\Omega)$ with a polynomial Q, $g \in C^1$, and $f \in C^\beta$ for some $0 < \beta \leq 1$. If*

$$N = \left\lceil m^{\frac{1}{2\beta+1}} \right\rceil, \quad n_1 = \ldots = n_m \geq \left\lceil m^{\frac{4\beta+17}{2\beta+1}} \right\rceil,$$

then using $\mathcal{H}$ with a constant R depending on d, Q, g, f, we have $\mathbb{E}\left[\left\|\pi_M f_{\hat{D},\mathcal{H}} - f_\rho\right\|^2_{L^2_{\rho_{\mathcal{U}}}}\right] = O\left(m^{-\frac{2\beta}{2\beta+1}} \log m\right)$.

Now we apply transformers to distribution regression.

With an input sequence $Q = [x_1 \cdots x_m]^T \in \mathbb{R}^{m\times d}$ of length m and feature dimension d, the single-head attention is defined as

$$\text{SoftmaxAttn}(x_i) = \sum_{j=1}^m \frac{\exp\left(\langle W_q x_i, W_k x_j\rangle/\sqrt{d_{in}}\right)}{\sum_{j'=1}^m \exp\left(\langle W_q x_i, W_k x_{j'}\rangle/\sqrt{d_{in}}\right)} W_v x_j,$$

where $W_q, W_k \in \mathbb{R}^{d_{in}\times d}, W_v \in \mathbb{R}^{d\times d}$ are parameter matrices.

We view the above as $\frac{1}{m}\sum_{j=1}^m k(x_i, x_j) f(x_j)$ with a data-dependent kernel k on $\Omega \times \Omega$ and $f : \Omega \to \mathbb{R}$ and introduce an attention operator $\text{attn}(\mu) = \int_\Omega k(\cdot, x) f(x) d\mu(x)$ and a transformer encoder as

$$c \cdot \sigma\left(A\left[\text{attn}(\mu)\right]|_{\mathbf{T}} + b\right),$$

where $\mathbf{T}$ is a set of n_2 points in Ω, $A \in \mathbb{R}^{n_1\times n_2}, b \in \mathbb{R}^{n_1}$ and $c \in \mathbb{R}^{n_1}$.

We generate the hypothesis space by bounding the parameters as $\|A\|_\infty \leq R m^{R\log m}, \|c\|_1 \leq R\sqrt{m}, \|b\|_\infty \leq R$ with $R > 0$.

Assume a Barron type condition

$$f_\rho = \Phi\left(\int_\Omega k(\cdot, x) f(x) d\mu(x)\right)$$

with a Mercer kernel k, f satisfying $c_f \leq |f(x)| \leq C_f$ with some $c_f > 0$, and the functional Φ having a representation

$$\Phi(g) = \int_{\mathcal{H}_k} e^{i\langle g,\omega\rangle_k} e^{i\theta(\omega)} F(d\omega), \qquad \|g\|_k \leq r$$

with some $r > 0, \theta : \mathcal{H}_k \to \mathbb{R}$ and a nonnegative function F satisfying $\int_{\mathcal{H}_k} \|\omega\|_k F(d\omega) < \infty$. Then we can achieve $\mathbb{E}\left[\left\|\pi_M f_{\hat{D},\mathcal{H}} - f_\rho\right\|^2_{L^2_{\rho_{\mathcal{U}}}}\right] = O\left(\frac{(\log m)^{d+2}}{\sqrt{m}}\right)$. See Theorem 5 in [4].

We can apply other structured deep neural networks [5, 6] to distribution regression.

REFERENCES

[1] Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1-40, 2016.

[2] A. Zweig and J. Bruna. A functional perspective on learning symmetric functions with neural networks. *International Conference on Machine Learning (2021), PMLR*, 13023-13032.

[3] Z. J. Shi, Z. Yu, and D. X. Zhou. Learning theory of distribution regression with neural networks. *Constructive Approximation*, (2025) 62:61-104.

[4] P. L. Liu and D. X. Zhou. Generalization analysis of transformers in distribution regression. *Neural Computation*, 37, 260-293, 2025.
[5] Y. F. Yang and D. X. Zhou. Nonparametric regression using over-parameterized shallow ReLU neural networks. *Journal of Machine Learning Research*, 25 (165), 1-35, 2024.
[6] D. X. Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48, 787-794, 2020.

SCHOOL OF MATHEMATICS AND STATISTICS, UNIVERSITY OF SYDNEY, AUSTRALIA
Email address: dingxuan.zhou@sydney.edu.au